# Multiple Imputation with Angular Covariates

## Imputing Incomplete Angular Data with Projected Normal Regression

Benjamin Stockton

University of Connecticut, Storrs, CT

6/5/23

# Table of contents I

# Table of Contents

# Overview

- Directional data consist of angles $\theta \in [0, 2\pi)$ but $0 \equiv 2\pi$, so standard inline methods are invalid
- Incomplete data arise in nearly every context data are collected including directional settings
- Incomplete angular data has only been addressed in a few, limited applications [16, 11, 10, 9, 14]
- We propose a novel application of the *projected normal regression* for imputing incomplete angular data

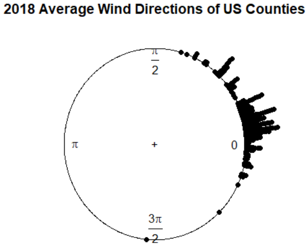


Figure 1: 2018 Average Wind Directions in US Counties

# Table of Contents

- Distributions on the circle can be intrinsic or arise out of transformation
- von Mises $M(\mu, \kappa)$ and Wrapped Normal $WN(\mu, \sigma)$ are unimodal and symmetric circular analog to the inline normal distribution [13]
- Projected Normal $PN(\mu, \Sigma)$ can be uni- or bi-modal and skewed or symmetric and is based on a latent bivariate normal distribution [17]

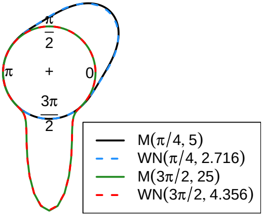


Figure 2: Examples of the von Mises and Wrapped Normal densities.

# Multiple Imputation

- Let $\mathbf{X} = (X_1, \ldots, X_p, Y)$ be complete and $\Theta = (\theta, \cos\theta, \sin\theta)$ be partially observed and $\epsilon \sim N_n(0, \sigma^2 I_n)$. We are interested in the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} \cos\theta + \beta_{p+2} \sin\theta + \epsilon$$

## Multiple Imputation Procedure [18] :

1. **Impute** - Use an imputation method $g$ to impute $\Theta_{mis}$ by $\dot{\theta} = g(\mathbf{X}, \Theta_{obs})$ or $(\cos\dot{\theta}, \sin\dot{\theta})' = g(\mathbf{X}, \Theta_{obs})$ $M$ times to create $M$ completed data sets

# Multiple Imputation

- Let $\mathbf{X} = (X_1, \ldots, X_p, Y)$ be complete and $\Theta = (\theta, \cos\theta, \sin\theta)$ be partially observed and $\epsilon \sim N_n(0, \sigma^2 I_n)$. We are interested in the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} \cos\theta + \beta_{p+2} \sin\theta + \epsilon$$

## Multiple Imputation Procedure [18] :

1. **Impute** - Use an imputation method $g$ to impute $\Theta_{mis}$ by $\dot{\theta} = g(\mathbf{X}, \Theta_{obs})$ or $(\cos\dot{\theta}, \sin\dot{\theta})' = g(\mathbf{X}, \Theta_{obs})$ $M$ times to create $M$ completed data sets
2. **Analyze** - For each completed data $(\mathbf{X}, \Theta_{obs}, \Theta_{mis}^{(m)})$ for $m = 1, \ldots, M$, estimate $\beta$ using least squares to collect $\hat{\beta}^{(m)}$ and $U^{(m)} = se(\hat{\beta}^{(m)})$

# Multiple Imputation

- Let $\mathbf{X} = (X_1, \ldots, X_p, Y)$ be complete and $\Theta = (\theta, \cos\theta, \sin\theta)$ be partially observed and $\epsilon \sim N_n(0, \sigma^2 I_n)$. We are interested in the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1}\cos\theta + \beta_{p+2}\sin\theta + \epsilon$$

## Multiple Imputation Procedure [18] :

1. **Impute** - Use an imputation method $g$ to impute $\Theta_{mis}$ by $\dot{\theta} = g(\mathbf{X}, \Theta_{obs})$ or $(\cos\dot{\theta}, \sin\dot{\theta})' = g(\mathbf{X}, \Theta_{obs})$ $M$ times to create $M$ completed data sets
2. **Analyze** - For each completed data $(\mathbf{X}, \Theta_{obs}, \Theta_{mis}^{(m)})$ for $m = 1, \ldots, M$, estimate $\beta$ using least squares to collect $\hat{\beta}^{(m)}$ and $U^{(m)} = se(\hat{\beta}^{(m)})$
3. **Combine** - Apply **Rubin's rules** to get point estimate $\bar{Q}$ and total variance $T$

# Table of Contents

# Inline Imputation Methods

- Let $\omega$ be one of $\{\theta, \cos\theta, \sin\theta\}$ depending on the imputation procedure with $\mathbf{X} = (1, Y, X_1, \ldots, X_p)$
- Assume only $\omega$ is incomplete

# Inline Imputation Methods

- Let $\omega$ be one of $\{\theta, \cos\theta, \sin\theta\}$ depending on the imputation procedure with $\mathbf{X} = (1, Y, X_1, \ldots, X_p)$
- Assume only $\omega$ is incomplete

## Linear Regression

- Use the observed data to fit the model $\omega_i = \mathbf{x}_i'\gamma + \eta_i$ for $i \in \{i : R_i = 1\}$; $\eta_i \overset{iid}{\sim} N(0, \tau^2)$
- Impute $\omega_i$ by drawing from the posterior predictive of the regression model given the predictors $\mathbf{x}_i$ for $i \in \{i : R_i = 0\}$

# Inline Imputation Methods

- Let $\omega$ be one of $\{\theta, \cos\theta, \sin\theta\}$ depending on the imputation procedure with $\mathbf{X} = (1, Y, X_1, \ldots, X_p)$
- Assume only $\omega$ is incomplete

## Linear Regression

- Use the observed data to fit the model $\omega_i = \mathbf{x}_i'\gamma + \eta_i$ for $i \in \{i : R_i = 1\}$; $\eta_i \overset{iid}{\sim} N(0, \tau^2)$
- Impute $\omega_i$ by drawing from the posterior predictive of the regression model given the predictors $\mathbf{x}_i$ for $i \in \{i : R_i = 0\}$

## Predictive Mean Matching

- Fit a Bayesian regression with a weakly informative prior and use draws from the posterior predictive to create a neighborhood set to draw donor points to serve as imputations

# Simulated Data Example

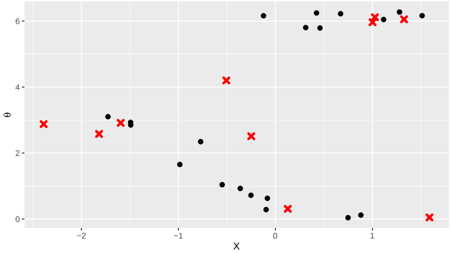


Figure 3: Viewed inline

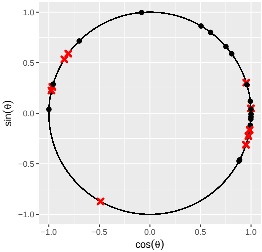


Figure 4: Viewed on the unit circle

Simulated data

# Just Another Variable Imputation

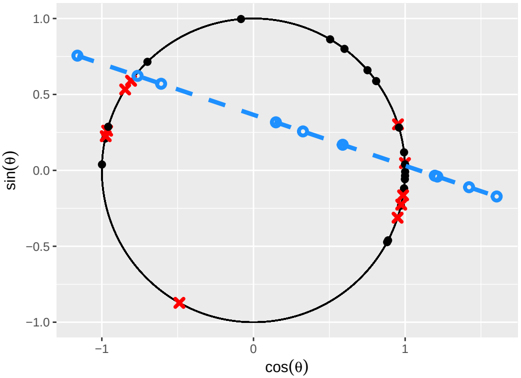


Figure 5: Impute on transformed angular data

# Passive Imputation

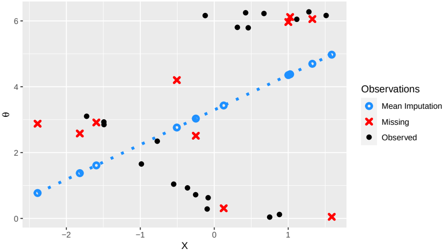


Figure 6: Imputed Inline. . .

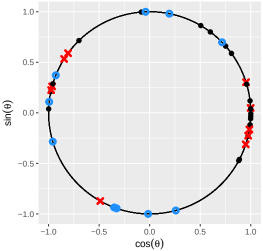


Figure 7: Transformed to cartesian coordinates

# von Mises Imputation

**GLIM-like Regression model**

- Let $\theta_i \sim M(\mu_i, \kappa)$ where $\mu_i = \mu_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\beta)$, $\beta \in \mathbb{R}^p$, and $\kappa > 0$ [15, 6]
  - Implemented by the `circglmbayes` and `brms` packages [3]
- Informative Normal prior on $\beta_j \sim N(0, 1)$ that encourages the constraint $|\beta_j| < 1.5$
- Uninformative prior on $\mu_0, \kappa$

# von Mises Imputation

### GLIM-like Regression model

- Let $\theta_i \sim M(\mu_i, \kappa)$ where $\mu_i = \mu_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\beta)$, $\beta \in \mathbb{R}^p$, and $\kappa > 0$ [15, 6]
    - Implemented by the `circglmbayes` and `brms` packages [3]
- Informative Normal prior on $\beta_j \sim N(0, 1)$ that encourages the constraint $|\beta_j| < 1.5$
- Uninformative prior on $\mu_0, \kappa$

**von Mises Imputation**

- Fit the von Mises regression model using the observed data and obtain posterior draws $\dot{\beta}$, $\dot{\mu}$, and $\dot{\kappa}$

# von Mises Imputation

## GLIM-like Regression model

- Let $\theta_i \sim M(\mu_i, \kappa)$ where $\mu_i = \mu_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\beta)$, $\beta \in \mathbb{R}^p$, and $\kappa > 0$ [15, 6]
  - Implemented by the `circglmbayes` and `brms` packages [3]
- Informative Normal prior on $\beta_j \sim N(0, 1)$ that encourages the constraint $|\beta_j| < 1.5$
- Uninformative prior on $\mu_0, \kappa$

### von Mises Imputation

- Fit the von Mises regression model using the observed data and obtain posterior draws $\dot{\beta}$, $\dot{\mu}$, and $\dot{\kappa}$
- Sample $\dot{\theta}_i \sim M(\dot{\mu}_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\dot{\beta}), \dot{\kappa})$

# von Mises Imputation

**GLIM-like Regression model**

- Let $\theta_i \sim M(\mu_i, \kappa)$ where $\mu_i = \mu_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\beta)$, $\beta \in \mathbb{R}^p$, and $\kappa > 0$ [15, 6]
  - Implemented by the `circglmbayes` and `brms` packages [3]
- Informative Normal prior on $\beta_j \sim N(0, 1)$ that encourages the constraint $|\beta_j| < 1.5$
- Uninformative prior on $\mu_0, \kappa$

## von Mises Imputation

- Fit the von Mises regression model using the observed data and obtain posterior draws $\dot{\beta}$, $\dot{\mu}$, and $\dot{\kappa}$
- Sample $\dot{\theta}_i \sim M(\dot{\mu}_0 + 2\,\text{arctan}^*(\mathbf{x}_i'\dot{\beta}), \dot{\kappa})$
- Use the posterior predictive draws $\dot{\theta}_i$ to impute the missing $\theta_i$

# Projected Normal Imputation

- Let $\theta_i \sim PN_2(\mu_i, I_2)$ where $\mu_i = \mathbf{x}_i'\mathbf{B}$ and $\mathbf{B} = (\beta_1, \beta_2) \in \mathbb{R}^{(p+2)\times 2}$
- Normal prior on $\beta_{ij} \sim N(0, 10000)$ [16, 5]
  - Implemented by `bpnreg` package in R
- Assumes a symmetric and unimodal distribution for $\theta_i$ given $\mathbf{x}_i$

# Projected Normal Imputation

- Let $\theta_i \sim PN_2(\mu_i, I_2)$ where $\mu_i = \mathbf{x}_i'\mathbf{B}$ and $\mathbf{B} = (\beta_1, \beta_2) \in \mathbb{R}^{(p+2)\times 2}$
- Normal prior on $\beta_{ij} \sim N(0, 10000)$ [16, 5]
  - Implemented by `bpnreg` package in R
- Assumes a symmetric and unimodal distribution for $\theta_i$ given $\mathbf{x}_i$

## Projected Normal Imputation

- Fit the projected normal regression model using the observed data and obtain posterior draws $\dot{\mathbf{B}}$ of $\mathbf{B}$

# Projected Normal Imputation

- Let $\theta_i \sim PN_2(\mu_i, I_2)$ where $\mu_i = \mathbf{x}_i'\mathbf{B}$ and $\mathbf{B} = (\beta_1, \beta_2) \in \mathbb{R}^{(p+2)\times 2}$
- Normal prior on $\beta_{ij} \sim N(0, 10000)$ [16, 5]
  - Implemented by `bpnreg` package in R
- Assumes a symmetric and unimodal distribution for $\theta_i$ given $\mathbf{x}_i$

## Projected Normal Imputation

- Fit the projected normal regression model using the observed data and obtain posterior draws $\dot{\mathbf{B}}$ of $\mathbf{B}$
- Sample $\dot{Y}_i \sim N_2(\mathbf{x}_i'\dot{\mathbf{B}}, I_2)$ and normalize to get $U_i = Y_i/||Y_i||^{-1} = (\cos\dot{\theta}_i, \sin\dot{\theta}_i)'$ for $i \in \{i : R_i = 0\}$

# Projected Normal Imputation

- Let $\theta_i \sim PN_2(\mu_i, I_2)$ where $\mu_i = \mathbf{x}_i'\mathbf{B}$ and $\mathbf{B} = (\beta_1, \beta_2) \in \mathbb{R}^{(p+2)\times 2}$
- Normal prior on $\beta_{ij} \sim N(0, 10000)$ [16, 5]
  - Implemented by `bpnreg` package in R
- Assumes a symmetric and unimodal distribution for $\theta_i$ given $\mathbf{x}_i$

## Projected Normal Imputation

- Fit the projected normal regression model using the observed data and obtain posterior draws $\dot{\mathbf{B}}$ of $\mathbf{B}$
- Sample $\dot{Y}_i \sim N_2(\mathbf{x}_i'\dot{\mathbf{B}}, I_2)$ and normalize to get $U_i = Y_i/||Y_i||^{-1} = (\cos\dot{\theta}_i, \sin\dot{\theta}_i)'$ for $i \in \{i : R_i = 0\}$
- Use the posterior predictive draws $\dot{\theta}_i$ to impute the missing $\theta_i$

# Example Imputations

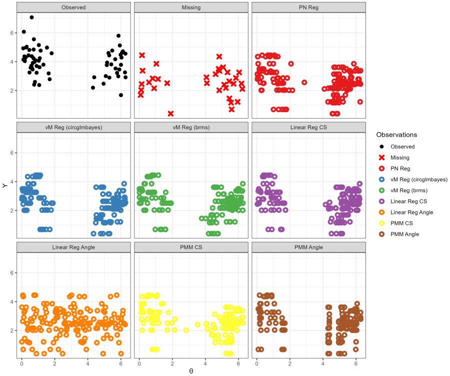


Figure 8: Imputations with angular data generated by von Mises regression. Viewed as angles/projected onto the unit circle.

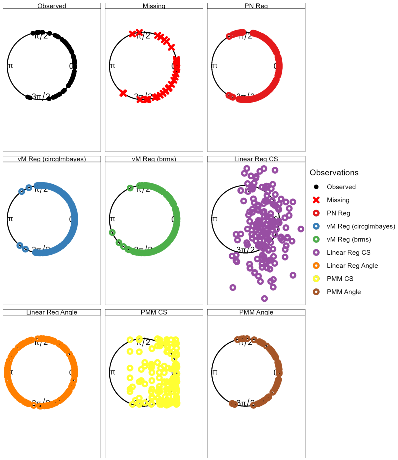


Figure 9: Imputations with angular data generated by von Mises regression. Viewed as coordinates in $\mathbb{R}^2$.

# Table of Contents

# Simulation Overview

How well do the projected normal imputations perform under various settings?

1. Different Sample Sizes - $N = 50; 100; 500; 1,000$
2. Different Missing Data Proportions - With $N = 100$, $p_{miss} = 0.1, 0.5, 0.9$
3. Different Data Generating Processes
    a. Low Concentration Projected Normal
    b. Skewed Projected Normal
    c. Bi-modal Projected Normal
    d. Projected Normal Regression
    e. von Mises Regression
    f. Wrapped Normal Regression

# Results I

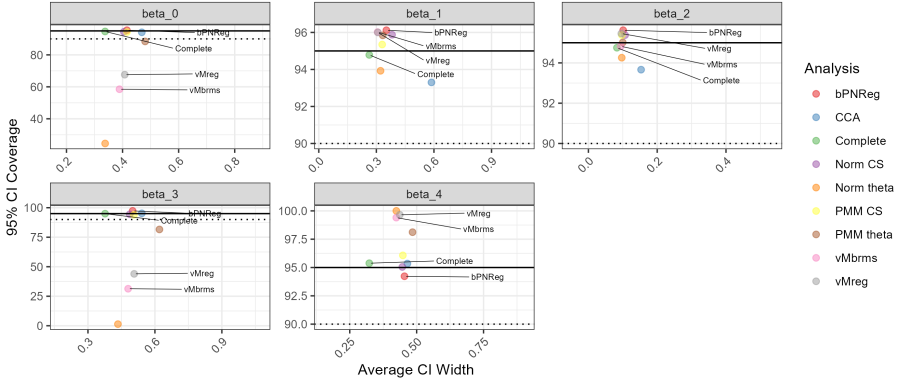


Figure 10: Labeled cases (1) the complete data, (2) projected normal (bPNReg), and (3) von Mises with brms (vMbrms), and (4) von Mises with circglmbayes (vMreg). Angle from $PN((1,0)', I_2)$.

# Results II

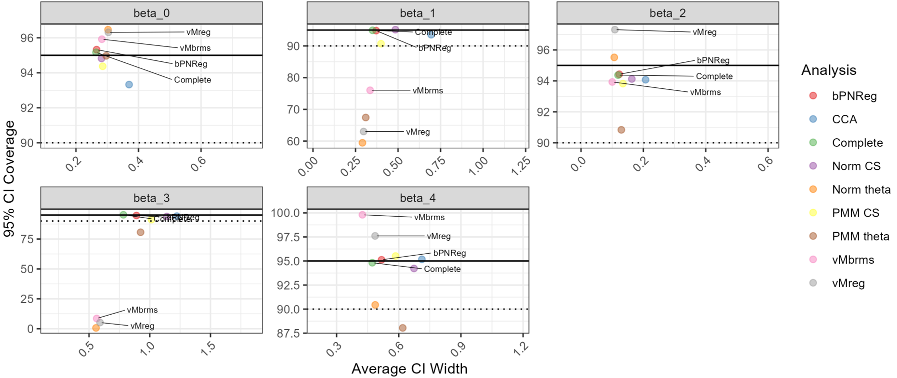


Figure 11: Labeled cases (1) the complete data, (2) projected normal (bPNReg), and (3) von Mises with brms (vMbrms), and (4) von Mises with circglmbayes (vMreg). Angle from Projected Normal Regression.

# Table of Contents

# Overview I

- Air pollution is strongly connected to asthma [7, 8, 12, 20, 21]
- Model the impact of air pollution on asthma accounting for meteorological conditions
- Data collected at the county-level and averaged over the year from CDC [4], EPA [19], and NOAA [1]
- Census data is used to weight the county observations by population [2]

# Overview II

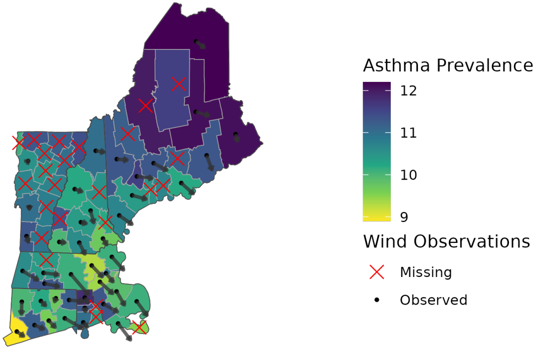


Figure 12: Map of New England counties asthma prevalence and average wind directions and speed.

# Analysis

- Impute using each of the previously discussed methods for the angular data
- Impute inline data with linear regression or predictive mean matching
- Include all variables collected for this step
- Mixed effects linear model (centered and scaled predictors)
    - Fixed effects for the meteorological and air pollution variables
    - Random intercepts for the state-level

$$
\begin{aligned}
Asthma_i = \beta_0 &+ \beta_1 AirPressure_i + \beta_2 RH_i + \beta_3 Temperature_i \\
&+ \beta_4 WindSpeed_i + \beta_5 \cos WindAngle_i + \beta_6 \sin WindAngle_i \\
&+ \beta_7 NO2_i + \beta_8 CO_i + \beta_9 SO2_i \\
&+ \beta_{10} PM2.5_i + \mathbf{Z}\eta + \epsilon_i
\end{aligned}
$$

# Results

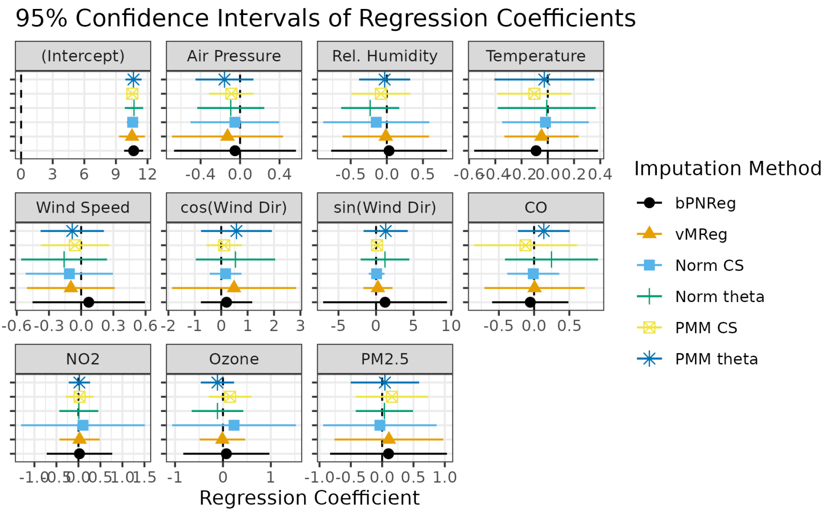


Figure 13: Coefficient estimates using different imputation strategies.

# Table of Contents

# What's Next?

- Imputation for spherical data
- Applying angular imputation models when the response or outcome is angular
- Incomplete data analysis for angular data in temporal or spatial settings

# Acknowledgements


- Research was supported by NSF AGEP-GRS supplement for Award #2015320.
- Thank you to my advisor Dr. Ofer Harel for support and assistance throughout the research process.


**Thank you!**

Questions?

# References I


[1] Local Climatological Data (LCD). http://www.ncei.noaa.gov/products/land-based-station/local-climatological-data, May 2021. Accessed: 5/19/2022.

[2] US Census Bureau. County Population Totals: 2010-2019. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html, 2021.

[3] Paul-Christian Bürkner. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80:1–28, August 2017. ISSN 1548-7660. doi: 10.18637/jss.v080.i01.

[4] Centers for Disease Control CDC and Prevention. National Environmental Public Health Tracking Network Data Explorer. https://ephtracking.cdc.gov/DataExplorer/?c=11. Accessed: 10/4/2022.

# References II


[5] Jolien Cremers. Bpnreg: Bayesian Projected Normal Regression Models for Circular Data, August 2021.

[6] N. I. Fisher and A. J. Lee. Regression Models for an Angular Response. *Biometrics*, 48(3):665–677, 1992. ISSN 0006-341X. doi: 10.2307/2532334.

[7] Jo Ann Glad, LuAnn Lynn Brink, Evelyn O. Talbott, Pei Chen Lee, Xiaohui Xu, Melissa Saul, and Judith Rager. The Relationship of Ambient Ozone and PM2.5 Levels and Asthma Emergency Department Visits: Possible Influence of Gender and Ethnicity. *Archives of Environmental & Occupational Health*, 67(2):103–108, April 2012. ISSN 1933-8244. doi: 10.1080/19338244.2011.598888.

# References III

[8] Jessie A. Gleason, Leonard Bielory, and Jerald A. Fagliano. Associations between ozone, PM2.5, and four pollen types on emergency department pediatric asthma events during the warm season in New Jersey: A case-crossover study. *Environmental Research*, 132:421–429, July 2014. ISSN 1096-0953. doi: 10.1016/j.envres.2014.03.035.

[9] Francesco Lagona and Marco Picone. Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics*, 39(5):927–945, May 2012. ISSN 0266-4763. doi: 10.1080/02664763.2011.626850.

[10] Francesco Lagona and Marco Picone. Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data. *Journal of Statistical Computation and Simulation*, 83(7):1223–1237, July 2013. ISSN 0094-9655. doi: 10.1080/00949655.2012.656642.

# References IV


[11] Francesco Lagona, Marco Picone, and Antonello Maruotti. A hidden Markov model for the analysis of cylindrical time series. *Environmetrics*, 26(8):534–544, 2015. ISSN 1099-095X. doi: 10.1002/env.2355.

[12] Mei Lin, Yue Chen, Richard T. Burnett, Paul J. Villeneuve, and Daniel Krewski. The influence of ambient coarse particulate matter on asthma hospitalization in children: Case-crossover and time-series analyses. *Environmental Health Perspectives*, 110(6):575–581, June 2002. ISSN 0091-6765. doi: 10.1289/ehp.02110575.

[13] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, Ltd, first edition, 1999. ISBN 0-471-95333-4. doi: 10.1002/9780470316979.

[14] Gianluca Mastrantonio, Giovanna Jona Lasinio, and Alan E. Gelfand. Spatio-temporal circular models with non-separable covariance structure. *TEST*, 25(2):331–350, June 2016. ISSN 1863-8260. doi: 10.1007/s11749-015-0458-y.

[15] Kees Mulder and Irene Klugkist. Bayesian estimation and hypothesis tests for a circular Generalized Linear Model. *Journal of Mathematical Psychology*, 80:4–14, October 2017. ISSN 0022-2496. doi: 10.1016/j.jmp.2017.07.001.

[16] Gabriel Nuñez-Antonio, Eduardo Gutiérrez-Peña, and Gabriel Escarela. A Bayesian regression model for circular data based on the projected normal distribution. *Statistical Modelling*, 11(3):185–201, June 2011. ISSN 1471-082X. doi: 10.1177/1471082X1001100301.

# References VI


[17] Brett Presnell, Scott P. Morrison, and Ramon C. Littell. Projected Multivariate Linear Models for Directional Data. *Journal of the American Statistical Association*, 93(443):1068–1077, September 1998. ISSN 0162-1459. doi: 10.1080/01621459.1998.10473768.

[18] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys / Wiley Series in Probability and Statistics*. Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley, New York, June 1987. ISBN 978-0-471-08705-2.

[19] OAR US Environmental Protection Agency. Air Quality System Data Mart [internet database]. https://www.epa.gov/outdoor-air-quality-data, August 2021.

# References VII


[20] C. P. Weisel, R. P. Cody, and P. J. Lioy. Relationship between summertime ambient ozone levels and emergency department visits for asthma in central New Jersey. *Environmental Health Perspectives*, 103 Suppl 2(Suppl 2):97–102, March 1995. ISSN 0091-6765. doi: 10.1289/ehp.95103s297.

[21] Adam M. Wilson, Cameron P. Wake, Tom Kelly, and Jeffrey C. Salloway. Air pollution, weather, and respiratory emergency room visits in two northern New England cities: An ecological time-series study. *Environmental Research*, 97(3):312–321, March 2005. ISSN 0013-9351. doi: 10.1016/j.envres.2004.07.010.